DOCUMENT-IDENTIFIER: US 20020059317 A1

TITLE: System and method for data management

--------- KWIC --------

Abstract Paragraph - ABTX (1):

An automated data management system and method for logging, processing, and reporting a large volume of data having **different file** types, stored on **different media,** and/or run by different operating systems, includes a first server processor for restoring a plurality of received data files, the data files being capable of being **different file** types; a file organizing/categorizing processor for organizing the received data files, based on a predetermined user list, into a source directory structure and a destination directory structure; a file logging processor for logging the received data files into a database formed by the source and destination directory structures and identifying a file type of the received data files; a de-duplicate processor for calculating a SHA value of the received data files to determine whether the received data files have duplicates and flagging duplicated data files in the database; an image conversion processor for converting the remaining data files into **image files,** respectively; and a second server processor for **exporting the image files**.

Summary of Invention Paragraph - BSTX (5):

[0003] To review and/or manipulate any of these data that are stored in **different file** types, **different media,** run by different operating systems, a customer often needs to open/close the corresponding different software programs, such as Word, WordPerfect, Excel, Email Outlook, etc. This is a very inefficient way of reviewing and manipulating the stored data. Further, one has to have these software programs and their updated versions to review and/or manipulate the stored data.

Summary of Invention Paragraph - BSTX (6):

[0004] In an area of litigation support, in particular, huge amount of documents and/or exhibits may have to be produced, organized, reviewed, reproduced, etc., for example, in merger and acquisition, intellectual property, anti-trust, and class action cases. The documents and/or exhibits may come from different locations in **different file** types. The existing methods of handling documents and/or exhibits include hand-coding or bar-coding. The hand-coding or bar-coding methods are not truly automated methods, and these methods are not efficient particularly in handling a volumetric amount of documents and/or exhibits.

Summary of Invention Paragraph - BSTX (12):
[0009] In one embodiment, a data management system in accordance with the principles of the present invention includes: a first server processor for restoring a plurality of received data files, the data files being capable of being **different file** types; a file organizing/categorizing processor for organizing the received data files, based on a predetermined user list, into a source directory structure and a destination directory structure; a file logging processor for logging the received data files into a database formed by the source and destination directory structures and identifying a file type of the received data files; a de-duplicate processor for calculating a SHA value of the received data files to determine whether the received data files have duplicates and flagging duplicated data files in the database; an image conversion processor for converting the remaining subset of de-duplicated data files into **image files,** respectively; and a second server processor for **exporting the image files**.

Summary of Invention Paragraph - BSTX (13):
[0010] Still in one embodiment, the **image files** are stored in the database to be viewed.

Summary of Invention Paragraph - BSTX (14):
[0011] Further in one embodiment, the **image files** converted from the data files are in a tiff format to be printed.

Summary of Invention Paragraph - BSTX (18):
[0015] Further in one embodiment, the data files having the same file type are converted into the **image files** together.

Summary of Invention Paragraph - BSTX (19):
[0016] Yet in one embodiment, the data management system includes a plurality of image conversion processors, each of the image conversion processors being capable of converting the data files having the same file type

into the corresponding **image files**.

Summary of Invention Paragraph - BSTX (22):
[0019] In one embodiment, the method in accordance with the principles of the present invention includes the steps of: restoring a plurality of received data files, the data files being capable of being **different file** types; organizing/categorizing the received data files, based on a predetermined user list, into a source directory structure and a destination directory structure; logging the received data files into a database formed by the source and destination directory structures and identifying a file type of the received data files; de-duplicating duplicates in the received data files by calculating a SHA value of the received data files to determine whether the received data files have duplicates and flagging duplicated data files in the database; converting the remaining data files into **image files, respectively; and exporting the image files**.

Summary of Invention Paragraph - BSTX (23):
[0020] Still in one embodiment, the method further includes the step of viewing the **image files** stored in the database.

Summary of Invention Paragraph - BSTX (24):
[0021] Further in one embodiment, the converting of the data files includes tiffing the data files into the corresponding **image files**.

Summary of Invention Paragraph - BSTX (27):
[0024] Still in one embodiment, the method includes parallel processing the steps of logging, converting, and **exporting** such that the data files are parallel-processed in a data file logging stage, an image conversion stage, and an image file output stage.

Summary of Invention Paragraph - BSTX (28):
[0025] Further in one embodiment, the converting of the data files includes converting the data files having the same file type into the **image files** together.

Summary of Invention Paragraph - BSTX (29):
[0026] Yet in one embodiment, the converting of the data files is processed by a plurality of image conversion processors, each of the image conversion

processors being capable of converting the data files having the same file type into the corresponding **image files**.

Summary of Invention Paragraph - BSTX (31):
[0028] One of the advantages of the present invention is that the data files are organized and processed in an efficient automated manner. The turn around time for generating a report containing the organized **image files** is substantially shortened.

Summary of Invention Paragraph - BSTX (34):
[0031] An additional advantage of the present invention is that the converted **image files** are organized such that it allows readily further processing of the data files.

Detail Description Paragraph - DETX (2):
[0040] The present invention discloses an efficient, automated data management system for logging, processing, and reporting a large volume of data
capable of being in different types, stored on **different media,** and/or run by a **different operating system**.

Detail Description Paragraph - DETX (4):
[0042] In FIG. 1, a plurality of data files N are **imported** into a data file input server processor 22. The data files are organized by a file organizing/categorizing processor 24 into a source directory structure and a destination directory structure. The data files are then logged into a file database 26 by a file logging processor 28. The file logging processor 28 identifies a file type of the data files and stores the file type information of the data files into the file database 26.

Detail Description Paragraph - DETX (5):
[0043] Also shown in FIG. 1, a de-duplicate processor 30 flags duplicates of the data files, i.e. de-duplicates the data files by creating a unique subset of data files by flagging duplicated files as such and storing this information the file database 26. Generally, the de-duplicate processor 30 calculates a SHA value of the received data files to determine whether the received data files have duplicates and flags duplicated data files in the file database 26. An image **conversion** processor 32 then **converts** the de-duplicated **data files into image files, and an image file** outputting server processor 34 **exports the**

image files.

Detail Description Paragraph - DETX (6):
[0044] The details of logging, de-duplicating, and converting the data files and outputting the corresponding **image files** are discussed in operation flows shown in FIGS. 2-6.

Detail Description Paragraph - DETX (7):
[0045] FIG. 2 illustrates an operation flow 36 of an exemplary data management method in accordance with the principles of the present invention. The operation 36 starts with an operation 38 of restoring a plurality of received data files. The data files can be of **different file** types. For example, the data files can be Word, JPEG, GIF, Bitmap, Excel, Access, Power Point, text, Adobe Acrobat, Paradox, ZIP files, etc. The data files are then organized, based on a predetermined user list, into a source directory structure and a destination directory structure in an operation 40. Next, in an operation 42, the received data files are logged into a file database formed by the source and destination directory structures. The operation 42 also identifies a file type of the received data files. Then, in an operation 44, the received data files are de-duplicated by calculating a SHA value of the received data files so as to determine whether the received data files have the same SHA value. If the data files have the same SHA value, then the data files are duplicates. If duplicates of the data files are found, they are flagged in the file database. The remaining de-duplicated data files are then converted into **image files** in an operation 46. Next, the converted **image files are exported** to a printer or a viewer, etc.

Detail Description Paragraph - DETX (8):
[0046] FIG. 3 illustrates an operation flow 50 of logging data files in accordance with the principles of the present invention, The logging data file operation 50 starts with an operation 52 of categorizing the received data files based on a predetermined user list and storing the data files in a data structure under a user directory. Then, the data files are categorized into email data files and user data files in an operation 54. For the email data files, an operation 56 determines whether there is an attachment to an email data file. If there is an attachment to an email data file, i.e. the "Yes" path, then the attachment is associated with the email data file in an operation 58 so that the **image files** of the attachment can be reviewed with the **image files** of the email data files. The attachment is then further categorized in the operation 54. If there is no attachment to an email data file, i.e. the "No" path, then the logging data file operation 50 ends. For

the user data files, on the other hand, the file type of the user data files is identified in an operation 60. For example, the data files having a Word format are distinguished from the data files having an Excel format. The data files having the same file type can be grouped and stored together in a database structure so that they can be processed together. Then, the logging data file operation 50 ends.

Detail Description Paragraph - DETX (10):

[0048] FIG. 5 illustrates an operation flow 72 of image conversion in accordance with the principles of the present invention. The image conversion operation 72 starts with an operation 74 of selecting a new file type to convert the data files under the selected file type into **image files**. Next, a new data file among the data files having the same file type is selected in an operation 76. Then, the selected **data file is converted into an image file** in an operation 78. Next, the image file is stored in the file database to be reviewed in an operation 80. If an operation 82 determines that there is another data file under the selected file type, then the operation flow 72 goes back to the operation 76 to select a new data file. If the operation 82 determines that there is no other data file under the selected file type, then the operation flow 72 goes to an operation 84 to determine whether there is another file type. If there is another file type in an operation 84, then the operation flow 72 goes to the operation 74 to select a new file type. If there is no other file type in the operation 84, the operation flow 72 is terminated.

Detail Description Paragraph - DETX (11):

[0049] FIG. 6 illustrates an operation flow 86 of outputting **image files** in accordance with the principles of the present invention. The outputting image file operation 86 starts with an operation 88 of identifying the **image files** that need to be processed in a report. Then, bates numbers for image file/slip sheets are generated in an operation 90. Next, slip sheets are generated to separate certain **image files** in an operation 92. Then, a review log is generated for further review and response to the report in an operation 94. Next, the report is outputted in a print format and/or an electronic viewer in an operation 96. Then, the operation flow 86 is terminated.

Detail Description Paragraph - DETX (18):

[0055] Meanwhile, an example of a destination directory and sub-directories for storing **image files** for an output report is created for Joe Smith's email as: Destination.backslash.Minneapolis.backslash.Email.backslash.9-12-88.backslash.Joe Smith.backslash..

Detail Description Paragraph - DETX (21):
[0057] The five phases of data processing include Logging/Extracting (Phase 1), Processing/Tiffing (Phase 2), Reporting**/Exporting** (Phase 3), Delivery/Printing (Phase 4), and Review/Second Print (Phase 5). The use of five phases allows one to control the quality and speed of data processing in each phase.

Detail Description Paragraph - DETX (27):
[0062] Figuring out if a data file is a duplicate or not. One way to achieve that is to use a SHA algorithm to determine a SHA value of a data file. SHA algorithm, i.e., Secure Hash Algorithm, was developed by the U.S. government to verify electronic transmissions of data between locations over fiber optic networks. The process analyzes and assigns a unique tag for each electronic document, based on the unique characteristics and patterns contained in the data. The SHA algorithm used in the present invention generates about 40 characters to identify a unique data file so as to determine whether there is a duplicate to the data file. If the two data files have the same SHA value, then the two data files are duplicates. Accordingly, the SHA value of a data file is compared to the existing SHA values in a database. If the SHA value has existed already, the data file is considered as a duplicate file. Accordingly, duplicated data files are flagged as duplicates and not converted into **image files**. Particularly in the litigation support area, removing duplicated data files saves review time by another person. Generally, this is no guarantee that two files are identical based solely on its file name, file dates, and file sizes. The method of generating SHA values for the data files in the present invention allows a mathematically certain process that prevents unique data from being overlooked and not processed.

Detail Description Paragraph - DETX (32):
[0067] In case of email PSTs (Personal Folders), **image files,** such as tiff images, of the email messages are generated, and any attachments found within the email are extracted.

Detail Description Paragraph - DETX (40):
[0074] Phase 2 is the step where **image files** (e.g. Tiff format files) of the logged data files are generated.

Detail Description Paragraph - DETX (48):
[0082] 3) **Converting the data or email file to an image file** and storing the

image file in the assigned user destination directory;

Detail Description Paragraph - DETX (58):
[0092] Once there are no more data files that need to be converted into image files, the particular user is considered done for Phase 2, ready for Phase 3. The master list of users is updated to indicate this.

Detail Description Paragraph - DETX (59):
Report and Export Step--Phase 3

Detail Description Paragraph - DETX (60):
[0093] Phase 3 is to generate ordered output for a customer or a print shop. Based on a master list of users, the directories and sub-directories that correspond to a particular user are selected for processing in Phase 3. The master list is updated to indicate that the particular user is in progress for Phase 3. Based on files tiffed up (i.e. the image files) in Phase 2, a report can be generated which contains a listing of all tiffed files. These image files are arranged in a hierarchy relationship. For example, email data files are arranged to be associated with their attachments.

Detail Description Paragraph - DETX (64):
[0097] Assigning a bates number to each page of the image files generated in sequential order. For example, page one of the email data file has a bates number of 100000. The first four-page attachment has abates number of
100001
to 100004. The second three-page attachment has a bates number of 100005
to
100007. In general, bates numbers are sequential for a particular user's data files. Each user may start at a pre-defined jump point of Bates. For example, user 1 starts at 1 and has 5000 pages, user 2 starts at 100000 and has 34000 pages, and user 3 starts at 200000 and has 345 pages. In this example, the jump point for Bates is 100000. Each user's data is separated by 100000. This allows us to assign bates numbers sequentially and still process more than one user at a time. It also provides that no two pages are going to have the same Bates Number. The information about the bates number is stored in a file database for running reports and a second report or print if desired (see below).

Detail Description Paragraph - DETX (80):

[0112] Shipping either a paper format of the processed documents, or the Tiffs being sent along with a log file that can be used to **import** into either an electronic viewer.

Detail Description Paragraph - DETX (86):

[0118] After a customer reviews the report generated, the customer may want to exclude and/or include some data files. The data files that are relevant are flagged. In this case, the data management system generates a new list of users and produces/prints only those **image files** that are flagged as relevant. A new set of sequential bates numbers are assigned. Slip sheets can be re-generated as described above if desired.

Detail Description Paragraph - DETX (87):

[0119] A process similar to Phase 3 is done here whereby only those documents that are marked as responsive are produced for print or **export**. A new set of bates numbers are assigned to the new subset of pages. All non-responsive documents are not considered for this re-print.

Claims Text - CLTX (2):

1. A data management system, comprising: a first server processor for restoring a plurality of received data files, the data files being capable of being **different file** types; a file organizing/categorizing processor for organizing the received data files, based on a predetermined user list, into a source directory structure and a destination directory structure; a file logging processor for logging the received data files into a database formed by the source and destination directory structures and identifying a file type of the received data files; a de-duplicate processor for calculating a SHA value of the received data files to determine whether the received data files have duplicates and flagging duplicated data files in the database; an image conversion processor for converting the remaining, de-duplicated, data files into **image files,** respectively; and a second server processor for **exporting the image files**.

Claims Text - CLTX (3):

2. The system of claim 1, wherein the **image files** are stored in the database to be viewed.

Claims Text - CLTX (4):

3. The system of claim 1, wherein the **image files** converted from the data

files are in a tiff format.

Claims Text - CLTX (10):
9. The system of claim 1, wherein the data files having the same file type are converted into the image files together.

Claims Text - CLTX (11):
10. The system of claim 1, wherein the data management system includes a plurality of image conversion processors, each of the image conversion processors being capable of converting the data files having the same file type into the corresponding image files.

Claims Text - CLTX (13):
12. A data management method, comprising the steps of: restoring a plurality of received data files, the data files being capable of being different file types; organizing/categorizing the received data files, based on a predetermined user list, into a source directory structure and a destination directory structure; logging the received data files into a database formed by the source and destination directory structures and identifying a file type of the received data files; de-duplicating duplicates in the received data files by calculating a SHA value of the received data files to determine whether the received data files have duplicates and flagging the duplicated data files in the database; converting the remaining data files into image files, respectively; and exporting the image files.

Claims Text - CLTX (14):
13. The method of claim 12, further comprising the step of viewing the image files stored in the database.

Claims Text - CLTX (15):
14. The method of claim 12, wherein the converting of the data files includes tiffing the data files into the corresponding image files.

Claims Text - CLTX (18):
17. The method of claim 12, further comprising the step of parallel processing the steps of logging, converting, and exporting such that the data files are parallel-processed in a data file logging stage, an image conversion stage, and an image file output stage.

Claims Text - CLTX (19):

18. The method of claim 12, wherein the converting of the data files includes converting the data files having the same file type into the **image files** together.

Claims Text - CLTX (20):

19. The method of claim 12, wherein the converting of the data files is processed by a plurality of image conversion processors, each of the image conversion processors being capable of converting the data files having the same file type into the corresponding **image files**.